Australian Bureau of Statistics

1352.0.55.120 - Research Paper: Using the EM Algorithm to Estimate the Parameters of the Fellegi-Sunter Model for Data Linking (Methodology Advisory Committee), Feb 2012

Latest ISSUE Released at 11:30 AM (CANBERRA TIME) 03/02/2012 First Issue

Summary

About this Release

Data linking is the act of linking two or more data files to bring together records which belong to the same individual. Data linking is performed at the Australian Bureau of Statistics (ABS) under the banner of the Census Data Enhancement Project, and involves linking Census data to administrative data sets. This data linking is done under the framework of the Fellegi–Sunter model. The parameters of this model need to be estimated for each linkage project. Previously the ABS has used training data to estimate these parameters, but there are limitations and drawbacks to this method. The use of the Expectation–Maximisation (EM) algorithm to estimate the parameters of the Fellegi–Sunter model is well established in the literature. This paper reviews and consolidates the existing research into using the EM algorithm for this purpose. It also documents the results of empirical work to investigate the behaviour of the algorithm on synthetic data sets where the true match status of the records is known.